CENTRO UNIVERSITÁRIO DA FEI

GUSTAVO MÜLLER NUNES

HISTOGRAMA DE ORIENTAÇÃO DE GRADIENTES PARA POSES DE MÃO EM UM AMBIENTE AUTOMOTIVO

São Bernardo do Campo
2014

A quem eu quero dedicar o texto.

AGRADECIMENTOS

Em construção.

“Em construção.”

.

RESUMO

Em construção.

ABSTRACT

Em construção.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Reconhecimento de gestos baseado em visão computacional é um assunto bastante pesquisado e já pode ser considerado popular nos dias de hoje, isto porque, a busca por mecanismos que tornem a interação entre homem e máquina mais intuitiva e natural é constante e vem aumentando com o lançamento de plataformas que auxiliam os desenvolvedores nos complexos algoritmos que envolvem essa área. O lançamento do Kinect, da Microsoft (MICROSOFT, 2014), e da plataforma de desenvolvimento da Intel, chamada Intel Perceptual Computing (INTEL, 2014), ambas com câmeras de profundidade, vem popularizando o desenvolvimento de aplicativos e revolucionando o jeito que interagimos com os jogos e computadores. Em fevereiro de 2013 a Microsoft anunciou que um terço dos consoles Xbox 360 vendidos até o momento tinham o sensor Kinect, totalizando uma venda de 24 milhões de sensores desde o lançamento do produto em novembro de 2010 (MICROSOFT, 2013).

O uso de câmeras em carros e caminhões também tem aumentando nos últimos anos. Sistemas de segurança capazes de verificar se o motorista esta saindo indevidamente da faixa, se o veículo esta em rota de colisão com algum outro automóvel, pedestre ou objeto e câmeras noturnas, que proporcionam ao motorista uma visão extra quando a estrada a frente esta escura, já são comuns em vários modelos de veículos. Em praticamente todos os modelos da Mercedes Benz, por exemplo, já é possível adquirir sistema de segurança desse tipo. Duas câmeras localizadas atrás do retrovisor e apontadas para a frente do veículo combinadas com um conjunto de radares fazem com que o motorista seja avisado caso aja risco de colisão, entre outras funcionalidades. Dependendo do caso o veículo pode atuar de forma autônoma e acionar os freios evitando uma colisão ou reduzindo a força do impacto (MERCEDES-BENZ, 2013; MERCEDES-BENZ, 2014).

Apesar do uso de câmeras para o lado externo do veículo já ser comum, ainda não é comum o uso dessas câmeras para interação do motorista com a grande quantidade de controles existentes nos carros. Sistemas de navegação, componentes de som e imagem como CD/DVD player, radio, televisão, celulares, computador de bordo, e ar condicionado são alguns exemplos de dispositivos que requerer uma constante interação e demandam uma grande quantidade de botões na região do console. Mesmo que os botões estejam agrupados em diferentes telas em um sistema multimédia, esse tipo de interação ainda exige uma grande atenção do motorista que precisa, na maioria dos casos, localizar visualmente o botão. Nem sempre também o uso dos botões ou da tela é confortável e ergométrico, podendo estar fora do alcance do motorista.

Uma maneira bastante natural de interagir com o veículo seria com voz e gestos. O sistema de voz já é bem comum hoje na maioria dos modelos de carros, mas são facilmente atrapalhados por barulhos internos e externo ao veículo e também interrompem o sistema de áudio, impossibilitando o seu uso caso, por exemplo, o sistema de viva voz estiver ativo. Por-

tanto um sistema de gestos pode complementar a interface existente, dando mais opção aos usuários na hora enviar comandos para o veículo.

Nesse sistema gestual de interação do motorista com o veículo, entende-se poses e gestos como sendo movimentos ou poses executados pela mão direta do motorista dentro do campo de visão de uma câmera instalada no teto do carro. O estudo desse trabalho, portanto, é focado na analise de um sistema capaz de caracterizar as poses de mão para que possam ser classificadas e usadas em um sistema de interface de usuário. Um sistema em tempo real capaz de reconhecer poses de mão e gestos que permita o motorista interagir com o veículo de forma intuitiva e eficaz. Na figura 1 tem-se um exemplo de como seria uma pose de mão aberta dentro de um ambiente automotivo.

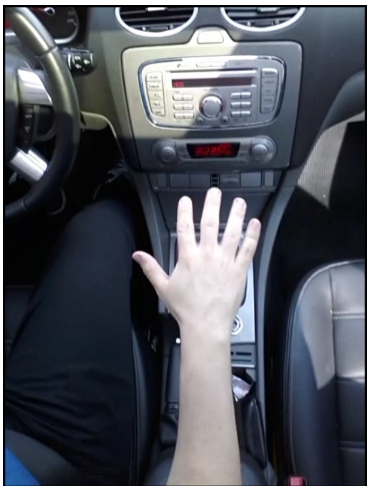


Figura 1 – Exemplo do uso de poses em um ambiente automotivo

1.1 Objetivo

O objetivo do trabalho é encontrar em uma imagem obtida no interior do veículo, do ponto de vista do teto e nas mais variadas condições de luminosidade, uma região de interesse com uma alta probabilidade de se ter uma pose de mão. A proposta é estudar o desempenho do histograma de orientações de gradientes (HOG - Histogram of Oriented Gradients) nessas condições e analisar o comportamento das variações dos parâmetros na aplicação e a influencia das variáveis do ambiente (mudança de veículo, de motorista e de luminosidade) no algoritmo. Um balanço entre performance e processamento deve ser levado em consideração, já que o trabalho computacional deve ser reduzido ao máximo para aplicações automotivas.

Um dos primeiros passos no processamento de uma imagem, com o objetivo de se identificar uma pose, é encontrar a região de interesse. Essa região é uma sub imagem onde a mão aparece com mais evidencia, eliminando as partes da cena que não são de interesse. Com o escopo reduzido, fica mais robusto aplicar outros algoritmos para a detecção da pose.

A escolha do HOG como descritor para as poses de mão se da por conta da sua grande robustez contra variações de luminosidade. O HOG foi proposto por Dalal e Triggs (2005) que encontraram o melhor conjunto de parâmetros para a representação de seres humanos em diversas situações e poses diferentes.

O descritor poderia ser usado de duas maneiras diferentes: como um pre classificador para encontrar as regiões mais prováveis de se ter uma mão e assim limitar a imagem em algumas regiões de interesse aonde um segundo algoritmo seria aplicado, nesse caso não seria trabalho do HOG dizer qual é a pose, mas sim se é uma mão ou não, ou no máximo classificar a pose em algum grupo de poses (como feito em (JIANG et al., 2012)). Mas o HOG poderia ser usado também para dizer qual pose é, sem a necessidade de nenhum algoritmo secundário. Visando que temos duas aplicações para o HOG, é possível que teremos duas configurações diferentes e portanto essas variações devem entrar no escopo desse trabalho.

Trabalha-se portanto com a hipótese de que o HOG pode ser parametrizado para encontrar poses de mão. E o conjunto original de parâmetros podem ser modificados para melhor de adequar à aplicação proposta.

1.2 Justificativa

A função principal do motorista deve ser sempre o controle do veículo. Distrações, como operar o rádio ou a central multimédia são exemplos constantes de cause de acidentes. Portanto apenas alguns poucos e curtos momentos podem ser usados para interagir com os comandos do veículo. Em estudos de usabilidade, o controle gestual provou ser mais intuitivo, efetivo (ZOBL et al., 2001) e distrair menos do que o uso habitual de botões (GEIGER et al., 2001). Por esse motivo, um estudo sobre técnicas para atingir esse objetivo é justificável.

As condições gerais dentro do automóvel inclui uma grande variação de iluminação, mudança de usuário (cor de pele, braço com ou sem vestimentas e vestimentas de cores e estampas diferentes) e fundos não uniformes. Além disso, a aceitação do usuário é um item bastante importante, portanto coisas como uma iluminação artificial visível, restrição de vestimentas e calibração extensiva não pode ser tolerados. Tento isso em mente, alguns critérios e requisitos para o sistema podem ser estabelecidos:

a) robustez contra ambientes ruidosos

b) iluminação invisível

c) independente de usuário

d) sem calibração ou treinamento pelo usuário

e) pequeno e compreensível conjunto de gestos

f) reação do sistema com o mínimo de latência

O estudo de Dalal e Triggs (2005) sobre histogramas de orientação de gradientes, aplicado à detecção de humanos variando cada parâmetro do cálculo dos histogramas e encontrando um conjunto de parâmetros que melhor servia para reconhecimento de humanos, virou referência para todos os estudos posteriores na área. Em seu texto ele diz que o uso de histogramas orientados tem muitos precursores (FREEMAN; ROTH, 1995; FREEMAN et al., 1996), mas que apenas atingiu a maturidade quando combinado com histogramas locais e normalização proposto pela Lowes Scale Invariant Feature Transformation (SIFT) (LOWE, 2004). A conclusão que ele chegou foi que usando histogramas de gradientes locais normalizados, similar ao SIFT, em um grade com sobreposição tem ótimos resultados para detecção de humanos, reduzindo falsos positivos em mais de uma ordem de magnitude comparado com Haar wavelets.

1.3 Metodologia

A primeira etapa do projeto é a captura das poses, a pesquisa literária mostra (ZOBL et al., 2004; AKYOL et al., 2000) que o uso de uma câmera infravermelha simples já é adequado para o problema, onde o ambiente é iluminado por infravermelho de curta distância (950nm). A câmera ainda possui um filtro de luz, permitindo apenas que a luz infravermelha seja capturada. Apesar de existir câmeras mais sofisticadas de alta resolução e tecnologias que permitem calcular a distância entre a câmera e o objeto, optou-se por usar a webcam simples por ser mais compatível com os padrões de mercado automotivo. No momento que esse texto foi escrito, as câmeras de profundidade, por exemplo, ainda possuem um preço proibitivo e a quantidade de processamento é bastante limitada em um ambiente embarcado.

As imagens de poses e os vídeos dos gestos serão obtidos em dois ambientes distintos. Primeiro em um ambiente controlado com fundo homogêneo de cor preta e em uma sala totalmente escura (essa base de dados será usada como referência para os algoritmos implementados). O outro será obtido no interior de um veículo, tanto de dia como de noite. A captura das imagens no interior do veículo é obtida variando tanto o motorista quanto o veículo. Também será usado outras bases de dados que não em veículos para comparar a performance do algoritmo nas mais diversas situações.

Próximo passo é a implementação do algoritmo HOG para depois variar os parâmetros e analisar com o melhor conjunto. Pode-se usar como referência a implementação feito pelo MATLAB, que usou os parâmetros de Dalal e Triggs (2005) em sua implementação e ainda permite um certo grau de parametrização.

Com relação às poses, vamos analisar a performance de 11 poses diferentes e analisar quais são as poses mais adequadas e que melhor se destacam para o uso em nossa aplicação. Para ter uma noção visual do quanto as poses são diferentes entre si, vamos reduzir as dimensões

do vetor de características gerado pelo algoritmo para apenas 3 dimensões. Usando o PCA podemos fazer essa redução e encontrar os 3 auto vetores e auto valores que melhor caracterizam nossas imagens e com isso conseguiremos um gráfico de três dimensões e ter uma perspectiva do quando os gestos estão agrupados entre si.

1.4 Organização da dissertação

Em construção.

2 REFERENCIAL TEÓRICO

O objetivo desse capítulo é referenciar as teorias que embasam o conteúdo desse trabalho bem como o estado da arte no uso do HOG.

2.1 Histograma orientado a gradientes

HOG (Histogram of oriented gradients) é um descritor computado a partir dos gradientes da imagem, podemos defini-lo com sendo uma informação estatística do gradiente e intensidade de uma área. Suas principais propriedades são a robustez para pequenas variações nos locais dos contornos, direções e variações significativas na iluminação e cor. Na figura 2 temos um resumo das principais etapas do cálculo feito para extrair o vetor de características.

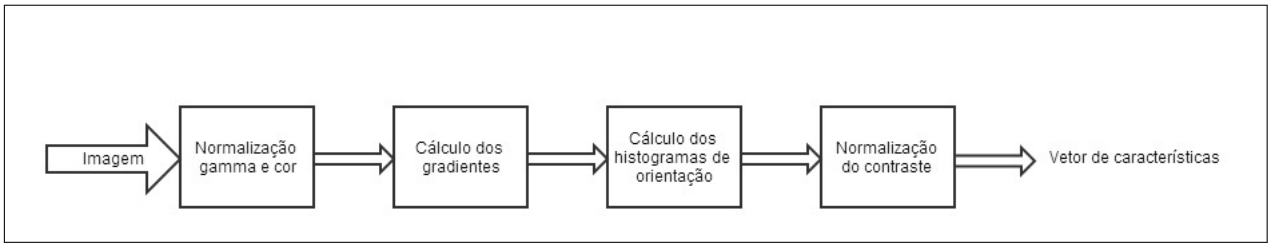


Figura 2 – Fluxo de cálculo para extrair o vetor de características

Para usar como referência, daqui pra frente vamos nos referir ao conjunto de parâmetros do HOG usado pelo Dalal e Triggs (2005) para detecção de pessoas como sendo o HOG original.

2.1.1 Normalização Gamma/Cor

Os pixels da imagem podem ser representados de diversas maneiras como escala de cinza, RGB e LAB. Uma normalização do gamma pode ainda ser aplicado. Apesar de imagens em tons de cinza apresentarem uma performance menor, essa será a opção de cor que iremos usar em nosso trabalho. O uso de câmeras infra vermelhas resulta na perda da informação de cor e portanto não teremos essa opção.

2.1.2 Gradientes

Um dos mais importantes processos no processamento de uma imagem é a sua segmentação. A segmentação consiste em subdividir a imagem em regiões ou objetos de interesse. O nível de segmentação depende do problema a ser resolvido e é comumente baseado em duas propriedades do valor da intensidade: descontinuidade e similaridade. A primeira consiste em particionar uma imagem baseado nas mudanças abruptas na intensidade, como por exemplo as bordas de um objeto. Já na segunda, é feito o agrupamento de uma região baseado em sua similaridade com outras partes da imagem, como cor ou nível de intensidade.

Gonzales (ano do livro) define borda como sendo um conjunto de pixeis conectados presente na fronteira entre duas regiões. E conclui que a magnitude da primeira derivada pode ser usada para detectar a borda em um ponto da imagem.

A derivada de primeira ordem de uma imagem digital pode ser aproximada no gradiente 2D. O gradiente de uma imagem $f(x, y)$ no ponto (x, y) e definido como um vetor

$$\nabla \mathbf{f}(x, y) = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f}{\partial x} \\ \dfrac{\partial f}{\partial y} \end{bmatrix} \tag{1}$$

cuja magnitude é definida como ∇f, onde

$$\nabla f = mag(\nabla \mathbf{f}) = \left[G_x^2 + G_y^2\right]^{1/2} \tag{2}$$

e a direção do vetor $\alpha(x, y)$ sendo definida como

$$\alpha(x, y) = tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{3}$$

onde o ângulo é medido em referência ao eixo x. A direção de uma borda no ponto (x, y) é perpendicular à direção do vetor gradiente no ponto.

O cálculo dessas derivadas podem ser implementados usando máscaras como o da figura 3. A máscara é aplicada em cada pixel da imagem e um novo valor é calculado conforme a equação 4.

$$R = w_1 z_1 + w_2 z_2 + w_3 z_3 + ... + w_9 z_9 = \sum_{i=1}^{9} w_i z_i \tag{4}$$

Nas figuras 4a e 4b temos dois exemplo das máscaras mais utilizadas para cálculo de gradiente. Na figura 5 podem ver o resultado das máscaras em uma imagem de uma pose de mão aberta feita por uma câmera infra vermelha.

No HOG original a máscara usada é uma máscara centrada 1-D [-1 0 1]. O gradiente é computado da seguinte maneira.

w_1	w_2	w_3
w_4	w_5	w_6
w_7	w_8	w_9

Figura 3 – Exemplo de máscara 3x3

-1	-1	-1	-1	0	1
0	0	0	-1	0	1
1	1	1	-1	0	1

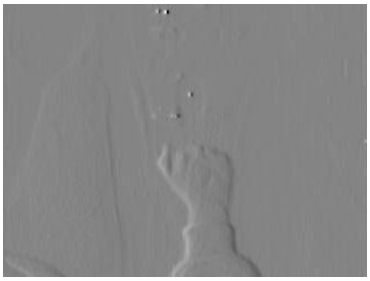
(a) Máscara Prewitt

-1	-2	-1	-1	0	1
0	0	0	-2	0	2
1	2	1	-1	0	1

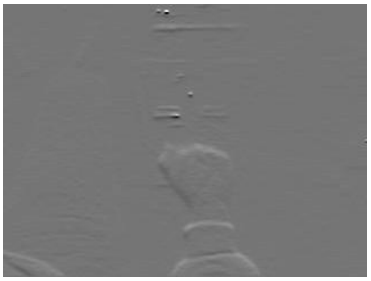
(b) Máscara Sobel]

Figura 4 – Exemplo de máscara de gradientes

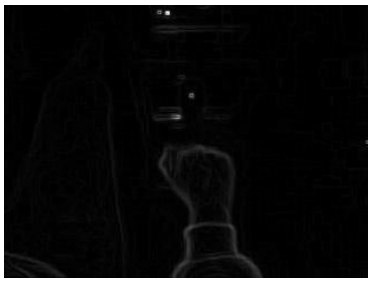
(a) Original



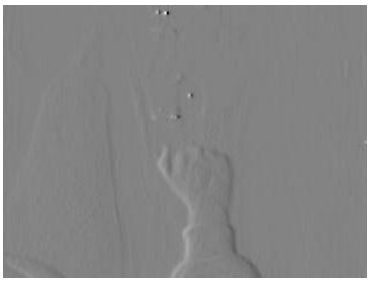
(b) Prewitt Gx



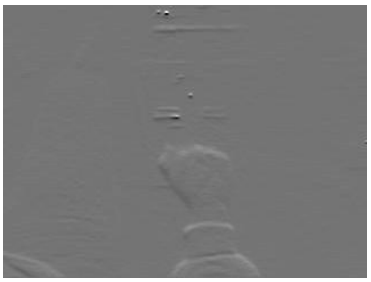
(c) Prewitt Gy



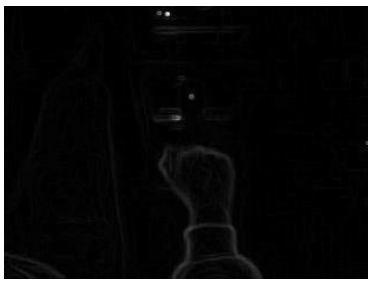
(d) Prewitt Gmag



(e) Sobel Gx



(f) Sobel Gy



(g) Sobel Gmag

Figura 5 – Gradientes

$$G_x(x, y) = f(x + 1, y) - f(x - 1, y)$$
$$G_y(x, y) = f(x, y + 1) - f(x, y - 1)$$

2.1.3 Cálculo dos histogramas

Depois dos cálculos do gradiente, a imagem é então dividida em pequenos retângulos (células). Para cada célula, um histograma é calculado. Esse histograma é a coleção dos ângulos dos vetores de gradiente de cada pixel que compõe a célula. Cada pixel apresenta um peso na construção do histograma das orientações das bordas. Esse peso pode ser em função do gradiente, do seu quadro ou da sua raiz.

Os ângulos podem ser agrupados variando de 0 à 360 graus ou de 0 à 180 graus.

No HOG original, as células tem tamanho 8x8, as orientações são ponderadas pela magnitude do vetor e uniformemente agrupas em 9 grupos de 0 a 180 graus.

2.1.4 Normalização em blocos

O tamanho do gradiente pode variar bastante por conta de variações como iluminação e contraste entre o fundo e o objeto de interesse. Portanto um importante passo para se obter um bom resultado na extração do vetor de características do objeto é a sua normalização.

Norma é uma função que atribui um tamanho de valor positivo e diferente de zero para um vetor em um espaço vetorial.

A função norma deve satisfazer algumas propriedades de escalabilidade e aditividade. Sendo um espaço vetorial V em um sub corpo F de números complexos, a norma em V é uma função $p :\rightarrow \mathbf{R}$ com as seguintes propriedades.

a) $p(a\mathbf{v}) = |a|p(\mathbf{v})$

b) $p(\mathbf{u} + \mathbf{v}) \leq p(\mathbf{u}) + p(\mathbf{v})$

c) Se $p(\mathbf{v}) = 0$ então $\mathbf{v}$ é o vetor zero.

Uma norma bastante usada é a norma euclidiana, que diz que em um espaço euclidiano R^n a norma será:

$$\|x\| := \sqrt{x_1^2 + ... + x_n^2}$$

Dos vários esquemas de normalização, a maioria é baseada no agrupamento de células em blocos maiores e normalizando o contraste de cada bloco separadamente. Além disso a uma sobreposição entre blocos para que as células de cada bloco possa contribuir nas componentes de normalização diversas vezes. Quatro esquemas foram testados por Dalal, sendo v um vetor não normalizado, $||v||_k$ sua $k - norm$ para $k = 1, 2$ e ϵ uma constante pequena temos:

a) L2-norm, $v \rightarrow \frac{v}{\sqrt{||v||_2^2+\epsilon^2}}$;

b) L2-Hys, L2-norm seguido por uma limitação nos valores máximos de v em 0.2 e renormalizando;

c) L1-norm, $v \rightarrow \frac{v}{||v||_1+\epsilon}$;

d) L1-sqrt, $v \rightarrow \sqrt{\frac{v}{\sqrt{||v||_2^2+\epsilon^2}}}$.

O HOG proposto por Dalal (**??**) possui a seguinte parametrização conforme tabela 1.

Cor	RGB sem correção de gamma
Gradiente	[-1, 0, 1] sem smoothing
Bins	9
Orientação	0 à 180
Tamanho do bloco	16x16 pixels
Tamanho da célula	8x8 pixels
Janela Gaussian	8 pixel
Normalização	L2-Hys
Janela de detecção	64x128

Tabela 1 – Parâmetros do HOG otimizado por Dalal

2.2 Estado da arte

Um dos precursores em extração de características da mão usando histograma de orientação de gradientes foi o laboratório da Mitsubishi que publicou um conjunto de artigos (FREEMAN; ROTH, 1995; FREEMAN et al., 1996) sobre o tema. Com o objetivo de identificar poses e gestos de mão para interfacear com aplicativos e jogos de computador, a abordagem foi aplicar o calculo de um histograma de orientações de gradiente na imagem convertendo a mesma em um vetor de características, que depois era comparado com um outro vetor de características de uma base de treinamento usando a distância euclidiana. Com apenas o calculo de um histograma, o sistema apresentava vetor de características muito parecidos para poses diferentes e exigia que a mão dominasse a área da imagem. Em imagens onde o tamanho da mão não era significativo (como a imagem de uma pessoa de corpo inteiro), a mudança da pose tinha pouco impacto no histograma.

Foi na elaboração do SIFT em (LOWE, 2004), que o uso da técnica do histograma de orientação de gradientes ficou genérico para o uso em diversas aplicações e acabou se tornou popular. O SIFT usa o vetor de gradiente de pontos chaves da imagem para gerar seu vetor de característica, mas a vantagem do método se da na normalização em blocos que aumenta o desempenho do algoritmo. Ele é conhecido por um algoritmo para detectar e descrever características locais da imagem. O algoritmo é patenteado nos Estados Unidos pela Universidade da Colúmbia Britânica.

Em (DALAL; TRIGGS, 2005) ...

Em (LI; CAO; XU, 2010) o HOG é utilizado para a detecção de ciclistas. No método proposto não foi feito overlap no cálculo dos histogramas, como uma maneira de melhorar o tempo de processamento, e amostragem piramidal é utilizada para extrair características globais em diferentes escalas. As imagens utilizadas são em tons de cinza e um filtro gaussiano é aplicado antes do cálculo dos HOGs (contrariando as orientações do Dalal e Triggs (2005)). O gradiente é calculado com máscara [-1 0 +1], os ângulos são calculados entre 0 e 180, e o histograma é dividido em 20 grupos de ângulos. A imagem é dividida em blocos de 16x16 sem divisão de células. O classificador utilizado é um SVM linear. Esse trabalho é interessante pois propõe um método para melhorar a velocidade do cálculo dos histogramas, o que pode ser útil para aplicações em tempo real embarcadas.

Um estudo comparando descritores locais, semi locais e globais é feito em (COLLUMEAU et al., 2011). O objetivo do trabalho é estudar qual seria o método mais adequado para descrever poses de mão em uma sala de cirurgia para que o médico possa enviar comandos para os aparelhos sem precisa encostar neles. Para descritores globais foi usado os momentos de Zernike (invariante em rotação, translação e escala) combinados com um classificador linear SVM. O HOG é usado como um descritor semi local e SIFT para locais. Apesar de não dar detalhes de como é feito os cálculos do HOG, o artigo mostra uma melhor performance do método. No melhor resultado encontrado, a taxa de reconhecimento do HOG foi de 87,66%, contra 73,32% do Zernike e 69,32% do SIFT.

Nesse artigo (LLORCA et al., 2011), o problema a ser resolvido era verificar, com o uso de uma câmera, se uma pessoa fez as seis diferentes poses de mão para o lavar correto das mãos. Primeiro as imagens são segmentadas por cor de pele e depois um estimador de posição do braço e da mão baseado em um filtro multi modal probabilístico é proposto. Um ROI é criado com o resultado do filtro anterior e então HOG é aplicado, usando como classificador dois SVM independentes. Uma para o HOG normal e outro para o HOF (Histogram of optical flow). Essa combinação espacial e temporal melhorou o desempenho do sistema aumentando a taxa de detecção.

Nesse artigo (KAWAHARA et al., 2012), é utilizado a coHOG (co-occurence HOG) para reconhecimento de navios em imagens ISAR. No coHOG os blocos são agrupados em pares, aumentado a robustez para imagens em diferentes ângulos e na oclusão de algumas partes do navio. Por outro lado, o coHOG tem uma alta dimensão. (melhorar)

A abordagem do artigo (JIANG et al., 2012) é criar um método para detectar as pontas dos dedos de uma mão de palma aberta com um câmera localizada em um óculos. A mão a ser detectada é a do próprio usuário do óculos, portanto as imagens serão de cima da mão. Um ângulo bastante semelhante ao do trabalho dessa pesquisa. Primeiramente a região da mão é encontrada usando o HOG como descritor e o SVM como classificador e posteriormente uma abordagem geométrica utilizando convex hull é aplicada para achar as pontas dos dedos. Os parâmetros utilizados para o cálculo do HOG foram células de 12x12 pixeis, com blocos de 2x2 células, ângulos variando de 0 a 180 graus agrupados em 9 regiões. Esse trabalho é uma boa referência que o HOG pode ser utilizado para encontrar uma região de interesse com alta probabilidade de se ter uma mão, mas infelizmente ele só abrange uma pose de mão.

Em (ZOBL et al., 2004) e em (AKYOL et al., 2000) temos um cenário automotivo idêntico ao proposto, onde imagens infra vermelhas de uma câmera instalada no teto do carro são capturadas e traduzidas em gestos e poses de mão. Em (ZOBL et al., 2004) o sistema proposto pelo artigo é capaz de reconhecer onze gestos e quatro poses. A imagem capturada com resolução 384x144 é primeiramente processada com uma combinação de subtração de fundo e threshold global. Em (AKYOL et al., 2000) é usado apenas um threshold global. A mão é considerada o maior objeto da cena. Depois da segmentação, um filtro para retirar o braço é aplicado e finalmente são calculados os momentos da imagem, para o cálculo da área e do centro de massa, e os momentos Hu. Usar os momentos Hu como vetor de características limita bastante a aplicação pois sua pose é representada por apenas 7 dimensões, o que parece um tanto quanto insuficiente. E a aproximação de que a mão é o maior objeto da cena é bem irreal, pois podemos ver, na base de dados extraída nessa pesquisa, que constantemente a perna do motorista ou o painel do veículo são os objetos maiores da cena.

Em (CHENG; TRIVEDI, 2008) e em (PARADA-LOIRA; GONZALEZ-AGULLA; ALBA-CASTRO, 2014) temos também o uso de câmeras infra vermelha no teto do carro. O primeiro tem com o objetivo de discriminar quem está usando o painel de controles do carro, o motorista ou o passageiro, e assim adaptar os controles para aumentar a segurança. O motorista quando usa o sistema de multimédia, tem a opção de controles reduzida para evitar distrações. Nesse estudo a posição do ROI é fixo e dividido em uma grade de células 2x2, o histograma é calculado para cada célula com 8 bins variando de 0 à 360 graus, portanto formando um vetor de 32 dimensões. O tamanho do ROI também é analisando variando entre 140x80, 80x80 e 140x140. O sistema faz uso de um classificador SVM e possui uma taxa de 96.8% de acerto. Esse trabalho mostra uma alta taxa de acerto usando um vetor de características de apenas 32 dimensões, mostrando um bom potencial para aplicações de tempo real.

Já o segundo tem a proposta de identificar poses e gestos para comandar o sistema multimédia do veículo. Para isso é feito uma combinação de remoção de fundo para segmentação e geometria computacional para classificação da pose. A segmentação é o resultado de três algoritmos rodando em paralelo, o Edge-based Foreground-Background Model (EBM), Mixture of Gaussians background Model (BG_MoG) e Maximally Stable Extremal Regions segmenter

(MSER). Essa estratégia resulta em um modelo para remoção de fundo estático e dinâmico, permitindo uma pre calibração do fundo com imagens estáticas e também se adaptando a eventuais mudanças por conta de novos objetos, sombras e diferentes condições de luminosidade.

3 HISTOGRAMA DE ORIENTAÇÃO DE GRADIENTES PARA POSES DE MÃO EM UM AMBIENTE AUTOMOTIVO

Esse capítulo tem como objetivo descrever as etapas da pesquisa prática. Conforme figura 6 os principais passos do projeto são a construção da câmera infra vermelha, a geração da base de dados, a seleção das imagens que servirão de treinamento para o classificador SVM e depois o cálculo do HOG em todas as imagens variando os seus parâmetros e medindo o desempenho do algoritmo.

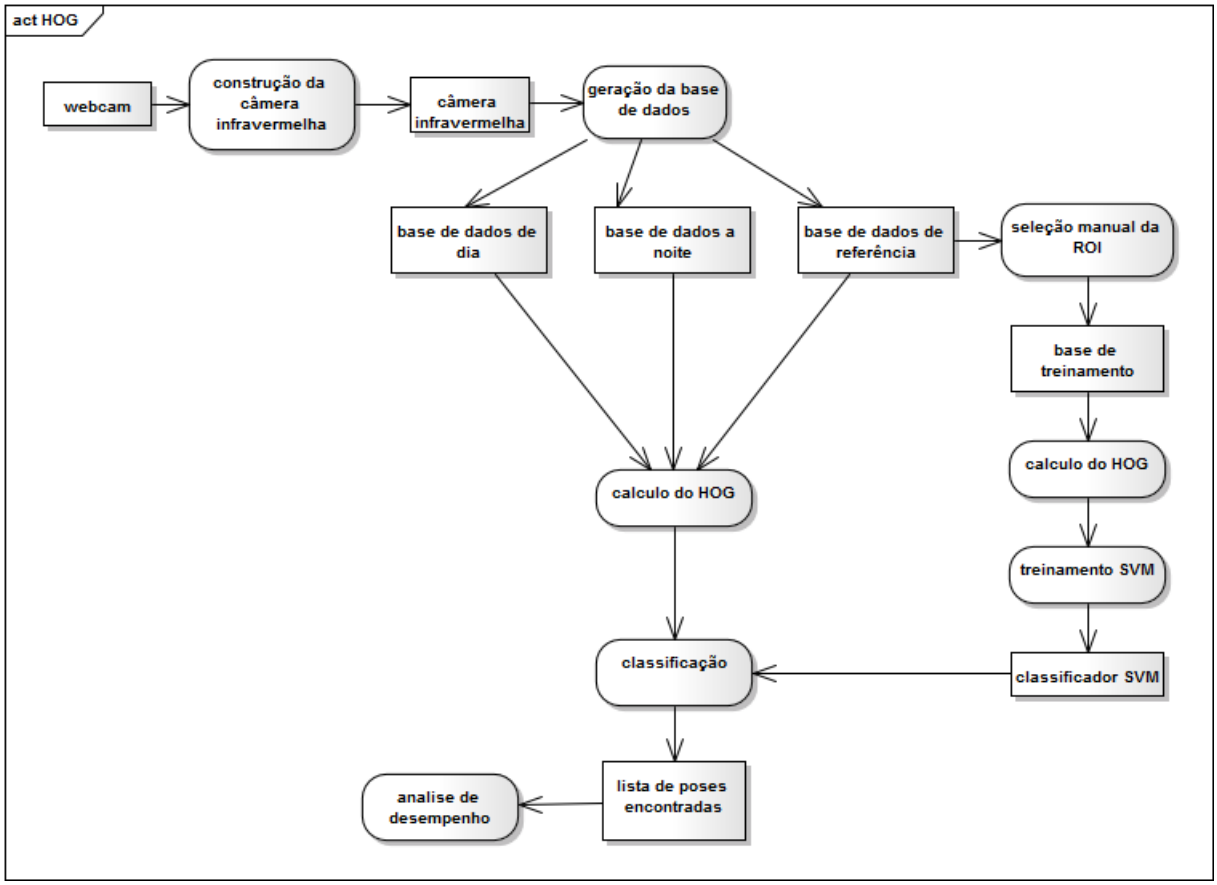


Figura 6 – Fluxo de trabalho da pesquisa

3.1 Construção da câmera infra vermelha

A câmera utilizada nessa aplicação tem que ser capaz de capturar imagens nas mais diversas condições de luminosidade. Temos o caso, por exemplo, de um dia de sol cuja intensidade de luz é bem alta. Até o ponto onde não há luz nenhuma. Nesses casos é necessário uma iluminação própria, mas ao mesmo tempo, não pode atrapalhar o motorista. Por isso, a ilumi-

nação infra vermelha é muito utilizada. O custo é baixo e não interfere em nada no ambiente. O maior contratempo desse tipo de iluminação é que se perde toda a informação de cor. Para gerar a base de dados para o nosso estudo, utilizamos uma câmera normal de mercado, modificada para receber a luz infra vermelha e colocamos LEDs de infra vermelho para fazer a iluminação.

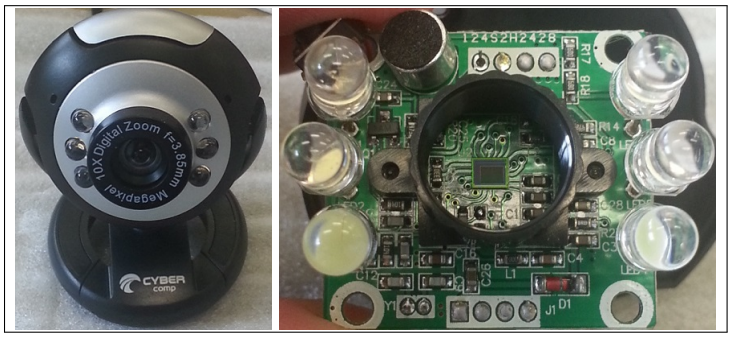


Figura 7 – Webcam utilizada na aquisição das imagens sem nenhuma modificação

Na figura 7 temos a câmera utilizada para a aquisição das imagens. Nesse momento a câmera ainda não foi modificada. Essa câmera portanto ainda possui um filtro de luz infra vermelha e os LEDs de iluminação são LEDs brancos.

A principal modificação a ser feita nesse tipo de câmera é retirar o filtro infra vermelho. Esse filtro é uma placa de vidro localizado atrás da lente. Na figura 8 temos uma foto das lentes ainda com o filtro e depois já com o filtro retirado. E preciso também substituir os LEDs atuais, que são LEDs brancos, para LEDs infra vermelho de 950nm.

Figura 8 – Lentes com o filtro infra vermelho localizado na parte traseira

3.2 Elaboração da base de dados

As bases de dados que serão usadas no trabalho precisam refletir as condições que encontramos em um ambiente automotivo. Por isso elaboramos um conjunto de banco de imagens que variam o fundo, o usuário, a iluminação e a vestimenta. A resolução das imagens será 320x240 e a janela terá o tamanho de 120x110 pixels.

O nosso fundo vai variar conforme o carro aonde as imagens foram coletadas. Como referência, temos também um conjunto de imagens com o fundo preto homogêneo. O usuário

será também modificado, variando sexo e cor de pele. A iluminação terá a captura diurna e noturna e a vestimenta varia por exemplo se o usuário esta usando blusa, relógio, etc.

Para o usuário temos a tabela 2 mostrando as principais características dos mesmos.

Usuário	Sexo	Cor de pele
1	Masculino	Branco
2	Masculino	Branco
3	Feminino	Branco
4	Masculino	Moreno
5	Feminino	Negra

Tabela 2 – Lista de usuários

A nossa base de referência será uma banco de imagens com o fundo preto homogêneo, o usuário 1 do sexo masculino sem nenhum tipo de vestimenta ou acessório e a iluminação apenas dos LEDs infra vermelho, ou seja, em um ambiente totalmente escuro. Na tabela 3 temos um resumo da parametrização dessa base e alguns exemplos das imagens.

Usuário	Usuário 1
Fundo	Preto homogêneo
Iluminação	Infra vermelha
Vestimenta	Nenhuma

Gestos

Tabela 3 – Parametrização da base de referência

Na tabela 4 temos um conjunto de imagens com o fundo do carro Ford Focus, iluminação com LEDs infra vermelhos e usuário 1 com uma blusa preta.

Usuário	Usuário 1
Fundo	Ford Focus
Iluminação	Infra vermelha
Vestimenta	Blusa preta

Gestos

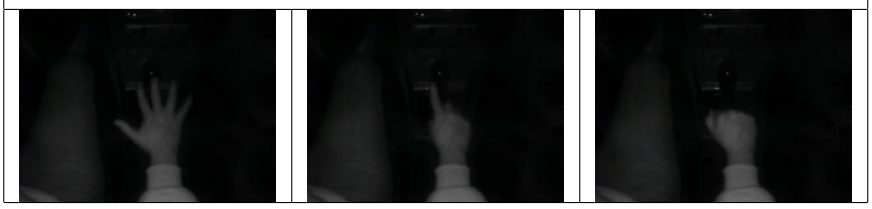


Tabela 4 – Parametrização do conjunto 2

Na tabela 5 temos um conjunto de imagens com o fundo do carro Ford Focus, iluminação com LEDs infra vermelhos e usuário 1 sem vestimentas.

Usuário	Usuário 1
Fundo	Ford Focus
Iluminação	Infra vermelha
Vestimenta	Nenhuma

Tabela 5 – Parametrização do conjunto 3

Na tabela 6 temos um conjunto de imagens com o fundo do carro Passat, iluminação com LEDs infra vermelhos e usuário 2 usando uma blusa verde. O interessante desse conjunto é a existência de um LED no painel que pode atrapalhar a segmentação da imagem.

Usuário	Usuário 2
Fundo	Passat
Iluminação	Infra vermelha
Vestimenta	Blusa verde

Tabela 6 – Parametrização do conjunto 4

3.3 Desenvolvimento da Pesquisa

Como vimos anteriormente na tabela 1, o HOG é calculado usando células de 8x8 pixeis, agrupas em blocos de 2x2 células. Portanto se aplicarmos o HOG com os parâmetros originais em uma imagem de 320x240, teremos 40x30 células. Cada célula contribui duas vezes para a formação do vetor de características por conta na sobreposição que existe na normalização em blocos, com exceção das bordas, que contribuem apenas uma vez. Portando teremos 40 + (40-2) x 30 + (30-2) células. Cada histograma tem 9 grupos de ângulos totalizando um vetor de 40.716 dimensões. Na figura 9 temos um exemplo visual do HOG. Cada histograma de cada célula é mostrado usando um diagrama de rosa. O tamanho de cada pétala do diagrama é ajustado para indicar a contribuição que aquela orientação representa no histograma da célula.

3.3.1 Implementação do HOG

Em construção.

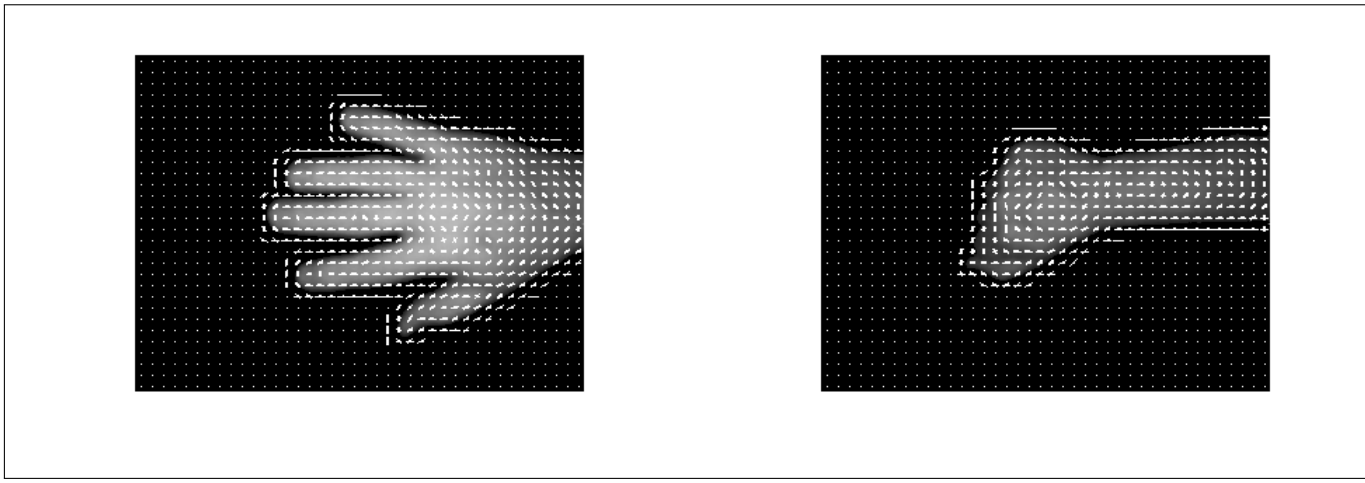


Figura 9 – Exemplos do cálculo do HOG com os parâmetros originais

3.3.2 Resultados

Para quantificar a performance dos classificadores que serão criados, optou-se pela geração de um gráfico do tipo DET (Detection Error Tradeoff). Esse tipo de gráfico é usado em classificações binárias medindo falso negativo vs. falso positivo em uma escala log-log. As curvas de gráficos do tipo DET tendem a ser mais lineares do que as curvas ROC (Receiver Operating Characteristics), facilitando a análise para pequenas variações. Os eixos serão: Taxa de Erro ($FalsoNeg/VerdadeiroPos + FalsoNeg$) versus Falso Positivo por Janela (FPPW do inglês False Positive per Window).

Primeiramente será criado um conjunto de classificadores para cada pose de mão que a pesquisa abrange. Esse conjunto é formado por 3 classificadores com foco na variação da luminosidade. Um para as imagens feitas durante o dia, um outro para as imagens infra vermelhas feitas à noite e um terceiro que seria genérico tanto para dia quanto para noite. O intuito é testar se a performance de um classificador específico para a luminosidade é melhor do que um classificador que abrange os dois tipos.

As imagens de treinamento serão geradas manualmente extraindo uma janela (110x120) com a pose de mão correspondente da base de referência.

Cada classificador será avaliado considerando imagens da pose versus imagens sem a pose e depois imagens da pose versus imagens de outras poses. Esse teste permite avaliar quais as poses são mais parecidas.

Considerando que temos 10 poses diferentes, teremos um total de 30 classificadores.

Depois um novo conjunto de classificadores será gerado (um para cada tipo de luminosidade) mas com treinamento de todas as poses com o objetivo de detectar mão independente da pose.

Esse conjunto de testes será repetido para cada parâmetro do cálculo do HOG conforme tabela **??**.

Número de células	8x8 até 240x240
Número de blocos	2x2 e 1x1
Tipo de normalização	L2-norm, L2-Hys, L1-norm , L1-sqrt
Agrupamento dos ângulos	9 a 36
Sinal dos ângulos	0-180 / 0-360

O tempo de cálculo será medido para cada parâmetro avaliado, para que depois se possa analisar o quão mais rápido o algoritmo se torna conforme seu desempenho cai.

4 DISCUSSÃO

O objetivo desse capítulo é discutir a relação entre a hipótese formulada no trabalho, a teoria existente sobre o assunto e a prática demonstrada no capítulo anterior.

5 CONCLUSÃO

Em construção.

REFERÊNCIAS

AKYOL, S. et al. Gesture control for use in automobiles. In: **MVA**. [S.l.: s.n.], 2000. p. 349–352.

CHENG, S. Y.; TRIVEDI, M. M. Real-time vision-based infotainment user determination for driver assistance. In: IEEE. **Intelligent Vehicles Symposium, 2008 IEEE**. [S.l.], 2008. p. 1–6.

COLLUMEAU, J.-F. et al. Hand-gesture recognition: comparative study of global, semi-local and local approaches. In: IEEE. **Image and Signal Processing and Analysis (ISPA), 2011 7th International Symposium on**. [S.l.], 2011. p. 247–252.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: IEEE. **Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on**. [S.l.], 2005. v. 1, p. 886–893.

FREEMAN, W. T.; ROTH, M. Orientation histograms for hand gesture recognition. In: **International Workshop on Automatic Face and Gesture Recognition**. [S.l.: s.n.], 1995. v. 12, p. 296–301.

FREEMAN, W. T. et al. Computer vision for computer games. In: IEEE. **Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on**. [S.l.], 1996. p. 100–105.

GEIGER, M. et al. Intermodal differences in distraction effects while controlling automotive user interfaces. In: **Proc. HCII**. [S.l.: s.n.], 2001. p. 263–267.

INTEL. **Intel Perceptual Computing**. 2014. [Online; acessado em 13-Agosto-2014]. Disponível em: <http://software.intel.com/en-us/vcsource/tools/perceptual-computing-sdk>.

JIANG, X.-H. et al. A robust method of fingertip detection in complex background. In: IEEE. **Machine Learning and Cybernetics (ICMLC), 2012 International Conference on**. [S.l.], 2012. v. 4, p. 1468–1473.

KAWAHARA, T. et al. Automatic ship recognition robust against aspect angle changes and occlusions. In: IEEE. **Radar Conference (RADAR), 2012 IEEE**. [S.l.], 2012. p. 0864–0869.

LI, T.; CAO, X.; XU, Y. An effective crossing cyclist detection on a moving vehicle. In: IEEE. **Intelligent Control and Automation (WCICA), 2010 8th World Congress on**. [S.l.], 2010. p. 368–372.

LLORCA, D. F. et al. A vision-based system for automatic hand washing quality assessment. **Machine Vision and Applications**, Springer, v. 22, n. 2, p. 219–234, 2011.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **International journal of computer vision**, Springer, v. 60, n. 2, p. 91–110, 2004.

MERCEDES-BENZ. **Mercedes-Benz TV: Active Lane Keeping Assist**. 2013. [Online; acessado em 13-Agosto-2014]. Disponível em: <https://www.youtube.com-/watch?v=OQkdvi55woA>.

MERCEDES-BENZ. **Mercedes Safety: Attention Assist, Pre-Safe and Distronic Plus**. 2014. [Online; acessado em 13-Agosto-2014]. Disponível em: <http://www.mbusa.com-/mercedes/benz/safety>.

MICROSOFT. **Microsoft News Center**. 2013. [Online; acessado em 13-Agosto-2014]. Disponível em: <http://www.microsoft.com/en-us/news/features/2013/feb13/02-11xbox-.aspx>.

MICROSOFT. **Kinect for Windows**. 2014. [Online; acessado em 13-Agosto-2014]. Disponível em: <http://www.microsoft.com/en-us/kinectforwindows/develop/>.

PARADA-LOIRA, F.; GONZALEZ-AGULLA, E.; ALBA-CASTRO, J. L. Hand gestures to control infotainment equipment in cars. In: IEEE. **Intelligent Vehicles Symposium Proceedings, 2014 IEEE**. [S.l.], 2014. p. 1–6.

ZOBL, M. et al. A usability study on hand gesture controlled operation of in-car devices. **Abridged Proceedings, HCI**, p. 5–10, 2001.

ZOBL, M. et al. Gesture components for natural interaction with in-car devices. In: **Gesture-Based Communication in Human-Computer Interaction**. [S.l.]: Springer, 2004. p. 448–459.